

# UC Irvine

## UC Irvine Previously Published Works

### Title

Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data

### Permalink

<https://escholarship.org/uc/item/4396n01s>

### Authors

Oprea, Alina  
Li, Zhou  
Yen, Ting-Fang  
[et al.](#)

### Publication Date

2015-06-01

### DOI

10.1109/dsn.2015.14

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data

Alina Oprea, Zhou Li, RSA Laboratories, Cambridge, MA, USA

Ting-Fang Yen, E8 Security, Palo Alto, CA, USA

Sang Chin, Draper Laboratory, Cambridge, MA, USA

Sumayah Alrwais, Indiana University, Bloomington, IN, USA

**Abstract**—Recent years have seen the rise of more sophisticated attacks including advanced persistent threats (APTs) [14], [33], [1], [2] which pose severe risks to organizations and governments by targeting confidential proprietary information. Additionally, new malware strains are appearing at a higher rate than ever before [29]. Since many of these malware are designed to evade existing security products, traditional defenses deployed by most enterprises today, e.g., anti-virus, firewalls, intrusion detection systems, often fail at detecting infections at an early stage.

We address the problem of detecting early-stage infection in an enterprise setting by proposing a new framework based on belief propagation inspired from graph theory. Belief propagation can be used either with “seeds” of compromised hosts or malicious domains (provided by the enterprise security operation center – SOC) or without any seeds. In the latter case we develop a detector of C&C communication particularly tailored to enterprises which can detect a stealthy compromise of only a single host communicating with the C&C server.

We demonstrate that our techniques perform well on detecting enterprise infections. We achieve high accuracy with low false detection and false negative rates on two months of anonymized DNS logs released by Los Alamos National Lab (LANL), which include APT infection attacks simulated by LANL domain experts. We also apply our algorithms to 38TB of real-world web proxy logs collected at the border of a large enterprise. Through careful manual investigation in collaboration with the enterprise SOC, we show that our techniques identified hundreds of malicious domains overlooked by state-of-the-art security products.

## I. INTRODUCTION

The cybersecurity landscape is evolving constantly. More sophisticated attacks including Advanced Persistent Threats (APTs) [14], [33], [1], [2] have emerged recently targeting organizations’ intellectual property, financial assets, and national security information. Well-funded attackers use advanced tools and manually orchestrate their campaigns to adapt to the victim’s environment and maintain low profiles of activity. Additionally there are also more malware than ever before. A whitepaper published by Panda Labs [29] found 30 million new malware strains in circulation in 2013 alone, at an average of 82,000 malware a day. Many of these are variants of known malware designed to evade existing security products, such that existing defenses, e.g., anti-virus, firewalls, intrusion detection systems, often fail at detecting infections at an early stage.

However, certain *infection patterns* still persist across malware variants and families due to the typical infection vectors used by attackers. For example, during the malware *delivery*

*stage*, victim hosts often visit several domains under the attacker’s control within a short period of time as a result of redirection techniques employed by attackers to protect their malicious infrastructures [36]. After delivery, backdoors are installed on the compromised machines to allow *footholds* into the targeted organization [26], where the machines initiate outbound connections regularly to a command-and-control server to receive instructions from the attacker. Malware communications commonly take place over HTTP/HTTPS, since web traffic is typically allowed by firewalls. More importantly, domains used in the same attack campaign are often related, meaning that they may share locality in either IP address space, time of access or set of hosts contacting them. These patterns of infections have been observed in advanced targeted attacks (e.g., APT1 group [26], Shady RAT [20], Mirage [12]), as well as botnet infections (e.g., Zeus, Citadel [13] and ZeroAccess [23]).

In this work, we leverage these observations to detect early-stage malware infections in enterprise networks. Our focus on enterprises stems from their unique point of view and new challenges present in those networks. As the victims of many cyber attacks, enterprises are pressed on detecting infections early to prevent further damage. However, the amount of network traffic generated by a large enterprise can be terabytes per day, requiring extremely efficient analysis methods to maintain a reasonable detection time. Yet, even though there is much data, it is of limited scope — only containing traffic as observed from that enterprise, making prior approaches using ISP-level data (e.g., [9], [6]) inapplicable.

We propose a graph-theoretic framework based on belief propagation [31] to identify *small communities* of related domains that are indicative of early-stage malware infections. We first restrict our attention to traffic destined to *rare destinations*. These are “new” domains, not visited before by any host in the organization within an observation window (and thus more likely to be associated with suspicious activity), and contacted by a small number of internal hosts (since we expect the initial infection to be small). In each iteration of our belief propagation algorithm, the rare domains are scored according to several features and similarity with domains detected in previous iterations.

Our algorithm can be applied either with “hints” (starting from “seeds” of known compromised hosts or domains), or without (when no information about compromised hosts or domains is given). In the first case, seeds can be obtained

from commercial blacklists and external intelligence sources containing Indicators of Compromise (IOCs) that the enterprise security operations center (SOC) has access to. Currently, SOC security analysts manually investigate incidents starting from IOCs, and we aim here to facilitate this process. In the latter case, our method first identifies automated, regular C&C communications that can then be used as seeds. Our C&C detector leverages unique properties of enterprise networks (e.g., popularity of user-agent strings and web referer information in the HTTP traffic) as well as features utilized in previous works (e.g., timing patterns, domain age and registration validity) and it can detect a single compromised host.

We demonstrate the effectiveness of our techniques on two different datasets, one containing DNS records and the other web proxy logs. The first consists of two months of anonymized DNS records from Los Alamos National Lab (LANL) in early 2013 amounting to 1.15TB. This dataset also includes 20 independent APT-like infection attacks simulated by LANL domain experts and was released along with a challenge problem to the community (*APT Infection Discovery using DNS Data [15]*) requesting methods to detect compromised internal hosts and external domains in the simulated attacks. The challenge included “hints” of varying details (e.g., one or multiple known compromised hosts), as well as answers for validation. Our techniques proved effective at detecting the LANL simulated attacks achieving an overall 98.33% true detection rate, at the cost of 1.67% false detection rate and 6.25% false negative rate.

Our second dataset contains two months of web proxy logs collected from a large enterprise in early 2014. Detecting malicious infections in this dataset proved to be more challenging due to its large scale (38TB of data), inconsistent information (host IP addresses are dynamically assigned using DHCP in most cases), and also the lack of ground truth. Through careful manual analysis in collaboration with the enterprise SOC, we identify hundreds of malicious domains not previously detected by state-of-the-art security products deployed on the enterprise network. Interestingly, a large number of these (98 distinct domains) are entirely new discoveries, not reported yet by VirusTotal several months after we detected them. This demonstrates the ability of our techniques to detect entirely new, previously unknown attacks.

To summarize our main contributions in the paper are:

**Belief propagation framework for detecting enterprise infection.** We develop a graph-theoretic framework based on belief propagation for detection of early-stage enterprise infections. Given “seed” hosts or domains, we automatically infer other compromised hosts and related malicious domains likely part of the same campaign. Our approach uniquely leverages relationships among domains contacted in multiple stages of the infection process.

**Detector of C&C communication in enterprise.** By exploiting novel enterprise-specific features and combining them with features used in previous work, we build a detector of C&C communication tailored to an enterprise setting. Domains labeled as potential C&C can be seeded in the belief propagation algorithm to detect other related domains.

**Solve the LANL challenge.** We apply the belief propagation algorithm to the LANL challenge and identify the malicious domains in the 20 simulated campaigns with high accuracy and low false detection and false negative rates.

**Evaluate on real-world data from large enterprise.** We apply our solution to a large dataset (38.41 TB) of web proxy logs collected at an enterprise’s network border. We identify hundreds of suspicious domains contacted by internal enterprise hosts which were not detected previously by state-of-the-art security products. Among the 375 domains detected in total over a month, 289 (accounting for 77.07%) are confirmed malicious or suspicious through careful manual investigation. While 191 (50.93%) are also reported by VirusTotal (but unknown to the enterprise of our study), we identify 98 (26.13%) that are entirely new discoveries (not reported by VirusTotal or the enterprise).

## II. PROBLEM STATEMENT

Our goal is to detect malware infection within an organization in early stages of the campaign. We describe below the characteristics of common enterprise infections (due to either targeted or opportunistic attacks), why existing solutions fail against such threats and the challenges we had to overcome for detecting them.

### A. Enterprise Infections

Common infection vectors for targeted attacks are social engineering [26] and compromise of legitimate sites [39]. In the case of social engineering, attackers craft legitimate-looking spear-phishing email addressed to several employees within the targeted organization including a malicious attachment or a hyperlink to a malicious file. Attack vectors employed by mainstream malware include spam emails, USB drives, and a variety of web-based attacks (e.g., drive by download, clickjacking, malvertising, etc.). Many of these attacks (both targeted and mainstream) have a common pattern during early-stage infection [26], [20], [12], [13]:

**Delivery stage:** During delivery, the victim machine gets the malicious payload, for example by an email attachment, or drive-by-download attack, etc. Many times, the first-stage malware is generic and needs to download additional malware (second stage) specifically crafted for the victim environment [38].

**Establishing foothold:** After delivery a backdoor is usually installed on the victim’s machine and the attacker establishes a foothold within the organization [26]. In almost all cases, backdoors initiate outbound connections to evade firewalls that block connections from outside the network. Most communications go through HTTP or HTTPS since these ports are allowed by most enterprise firewalls [34], [28].

**Command-and-control (C&C):** Typically, backdoors connect regularly to the command-and-control center operated by attackers to receive further instructions and allow attackers backdoor access into the victim environment [26], [13].

Based on these infection patterns, we extract several common characteristics of enterprise infections:

**Small scale:** Attackers are motivated to maintain stealthy operations to avoid detection by security products deployed within enterprise perimeter and we focus on detecting infections with small scale. In some cases a single host might get infected and communicate with the C&C center.

**Uncommon domains:** Attackers tend to use uncommon destinations for different stages of the campaign (e.g., delivery, C&C). To examine the popularity of domains used in attacks, we obtained a list of 14,915 IOCs reported between 2011 and 2014 from the SOC of a large enterprise. None of the indicators are among the Alexa top one million most popular domains [4]. Additionally, [26] points out that attackers use more frequently domain names rather than direct IP connections for their C&C communication so that they can dynamically flux the domains. Among the 14,915 IOCs in the enterprise list, the vast majority (13,232 or 88.71%) are domain names.

**HTTP/HTTPs Communication.** The communications between malware and C&C servers is typically done through HTTP or HTTPs since other ports are blocked by enterprise firewalls [34], [28].

**Communities of domains:** A compromised host usually contacts several malicious domains within a relatively short time interval. For instance, a user clicking on an embedded link in an email might visit the front-end attacker site, get redirected to a site hosting malicious payload and shortly after the backdoor is established will initiate the first connection to the C&C server. These domains form *small communities* exhibiting similarity in connection timing, set of hosts contacting them (if multiple hosts are infected in the same campaign) and sometimes proximity in IP address space [19], [26].

**Automated C&C communication:** Backdoors typically communicate with C&C servers on a regular basis to allow attackers access into the victim environment. In many publicized APT campaigns (e.g., NightDragon [11], Mirage [12], Shady RAT [20]) as well as botnet infections (e.g., Zeus, Citadel [13], ZeroAccess [23]), C&C communication occurs at regular time intervals (minutes or hours). We also examined malware samples provided by Mandiant on the APT1 group to test their communication patterns. Among 43 backdoor samples, only 4 exhibit randomized communication patterns while the remaining ones periodically communicate back to C&C servers (with small variation between connections).

## B. Current Defenses

Large enterprises deploy different security products (anti-virus, intrusion-detection, firewalls, etc.). Detection and response are largely the responsibility of the Security Operations Center (SOC). The SOC consists of a team of security analysts that monitors the network activities inside the enterprise network, often relying on commercial blacklists or external intelligence sources (i.e., IOCs) to identify malicious domains and infected hosts, followed by manual investigation to determine the scope of the threat. As IOCs are by no means complete, the investigation phase is particularly labor-intensive. It is this process that we aim to facilitate in this paper.

Systems to detect botnet communication or malicious domains have been extensively proposed in the literature. Some of them (e.g., [37], [22]) require malware samples for training. Unsupervised systems (e.g., [16], [18], [41]) typically require multiple synchronized hosts compromised by the same malware, and do not scale to large networks. ExecScent [28] identifies C&C domains from enterprise web traffic, though it requires malware samples as input to generate traffic templates. Our approach is also tailored to large enterprise networks like ExecScent, but does not require malware binaries, and can identify multiple related malicious domains used in the campaign (not only the C&C stage).

## C. Challenges

There were several challenges we had to overcome in the process of developing our detection methodology. First, security products deployed within the enterprise perimeter record large volumes of data daily. For example, the two datasets we used to evaluate our system are 1.15 TB and 38.14 TB, respectively. To perform efficient analysis, we describe in §IV-A a suite of techniques that reduce the data volume by an order of magnitude while retaining the valuable information about communication of internal hosts to external domains.

In practical situations, we have to deal with different log formats generated by a variety of deployed security products. We focus on general patterns of infections that is common in various types of network data (e.g., NetFlow, DNS logs, web proxies logs, full packet capture) and demonstrate the effectiveness of the algorithms on two distinct datasets (DNS logs in §V and web proxies logs in §VI). In the web proxies log dataset we enrich the set of features with context information available in HTTP connections.

Finally, while most existing detection systems focus on general malware, we aim to also tackle sophisticated infections which could be part of a targeted attack. APT attacks are extremely stealthy, maintain low profiles and adapt to the victim environment, leaving only small amount of evidence in the log data. We develop an algorithm for identifying suspicious C&C domains in §IV-C (even when contacted by a single host) requiring no prior knowledge of the malware sample. We use these domains as seeds to identify other related malicious domains and compromised hosts part of the same campaign through belief propagation. In the case when IOCs are provided (which is common in enterprise settings), they can be used as seeds in the same belief propagation algorithm. Our algorithms are unique in identifying relationships among domains used in different infection stages, e.g., delivery, C&C. The evaluation on LANL dataset shows that our approach is able to detect simulated APT infections and thus has the potential of detecting targeted attacks in the wild.

## III. METHODOLOGY

In this section, we provide an overview of our approach to detecting early-stage enterprise infection. Our system analyzes log data collected at the enterprise border on a regular basis (e.g., daily), maintains profiles of normal activity within the enterprise, and detects malware infections by exploiting the

relationship between suspicious external destinations used in different infection stages. We first describe the datasets used in this study, then introduce our main framework based on belief propagation, the features for detecting suspicious external destinations, and conclude with an overview of the system operation. Details on our techniques and features are given in the next section.

#### A. Datasets

The first dataset used in this study consists of anonymized DNS logs collected from the Los Alamos National Lab (LANL) internal network. It includes DNS queries initiated by internal hosts, DNS responses from the LANL DNS servers, timestamps of those events, and IP addresses of the sources and destination hosts. All of the IP addresses and domain names are anonymized (consistently — i.e., the same IP always maps to the same anonymized IP). In addition to real-world DNS logs, the dataset also includes simulated DNS traffic that is representative of those observed during the initial stage of stealthy, targeted attacks (i.e., APTs). A total of 20 distinct attack campaigns are included.

The second dataset *AC* consists of logs collected by web proxies that intercept HTTP/HTTPS communications at the border of a large enterprise network with over 100,000 hosts. The logs include the connection timestamp, IP addresses of the source and destination, full URL visited, and additional fields specific to HTTP communications (HTTP method, status code, user-agent string, web referer, etc.). As such, the *AC* dataset is much richer compared to the LANL dataset.

In addition to the web proxy logs, we also obtained access to a list of domain IOCs used by the SOC for detecting malicious activities, and compromised internal hosts communicating to those domains. The IOCs are gathered from external intelligence sources and commercial blacklists, and often serve as the initial point of investigation by SOC analysts. This data was collected at the same time as the web proxies logs.

All of our datasets span an interval of two months (in early 2013 for LANL, and early 2014 for *AC*) and are extremely large (1.15TB and 38.14TB, respectively), raising a number of scalability challenges. While the *AC* dataset is much richer in information, the main advantage of LANL is that it includes the simulated infection campaigns by LANL domain experts and thus provides a labeled dataset by which to validate our methods. Our main results, however, are from evaluations performed on the *AC* dataset in collaboration with the enterprise SOC.

In the analysis and results presented in the following sections, we focus on “rare” destinations in our datasets. Our insight is that popular websites (visited by a large user population) are better administered and less likely to be compromised, but connections to uncommon destinations may be indicative of suspicious behavior. More specifically, we are interested in external destinations that are:

- *New*: Not visited before by any internal hosts. The rationale is that attackers tend to use new domains under their control for at least some of the attack stages, and also that those malicious domains would not be contacted by benign hosts.

- *Unpopular*: Visited by a small number of internal hosts. The intuition is that attackers are likely to compromise only a few hosts during initial infection.

To identify the “new” destinations, we keep track of external destinations contacted by internal hosts over time. This “history” of destinations is initialized during a bootstrapping period (e.g., one month), and then updated incrementally daily.

These new and unpopular domains are called *rare destinations* and are the starting point of our detection. We found in the enterprise of our study that the number of rare destinations is on the order of 50,000 daily, and the challenge we face is identifying in this set potential malicious domains.

#### B. Modes of operation

Our detection method operates in two modes. In the first, called *SOC hints*, we use the incidents that the enterprise SOC investigated as starting points. Given either hosts or domains confirmed malicious by an analyst as seeds, our algorithm identifies other related malicious domains (likely part of the same campaign) and internal compromised hosts that were unknown previously.

The second is the *no-hint* mode, in which no known compromised hosts or malicious domains are available. In this mode, we develop a new C&C communication detector utilizing connection timing patterns from hosts to domains, domain registration information, and enterprise-specific features extracted from the web proxy logs. Compared to previous work, our C&C detector does not require malware samples for training, and can detect C&C communication when only one infected host in the enterprise communicates with the external C&C domain. Interestingly, the detected C&C domains and the hosts contacting them can be used to seed the same belief propagation algorithm and identify additional related suspicious domains and compromised hosts, i.e., serve as input to the *SOC hints* mode.

#### C. Belief propagation framework

We model the communication between internal hosts and external domains with a bipartite graph  $G = (V, E)$ , in which there are two types of vertices, hosts and domains. An edge is created between a host and a domain if the host contacts the domain at least once during the observation window (e.g., one day). We would like to label each domain as malicious or benign, and each host as compromised or clean, with high confidence.

Our main insight is to apply a graph theoretic technique called *belief propagation* [31], commonly used to determine the label of a node given prior knowledge about the node itself and information about its graph neighbors. The algorithm is based on iterative message-passing between a node and its neighbors until convergence or when a specified stopping condition is met. For our purposes, the messages (or “beliefs”) passed between nodes are an indication of suspicious activities observed during early-stage malware infections. In typical implementations, the entire graph is constructed in advance and “beliefs” are transmitted from every node to all its neighbors in each iteration. Since the graphs in our case are very large

(potentially tens of thousands of domains and hosts daily), we propose an *incremental* method of building the bipartite graph in which hosts and domains are added to the graph only when the confidence of their compromise is high.

We build a graph starting from the known compromised hosts (and optionally suspicious domains) given as seeds. In each iteration, we compute scores for those rare domains contacted by known compromised hosts, and label the domains with the highest scores as suspicious. The set of compromised hosts is also expanded to include other hosts contacting the newly labeled suspicious domains. The algorithm terminates when the score of the top-ranking domain is below a threshold, or when the maximum number of iterations is reached, and returns a list of labeled malicious domains ordered by suspiciousness level.

The score for a domain is computed based on 1) the degree to which the domain exhibits C&C-like behavior (described in §III-D), or 2) its similarity to labeled suspicious domains from previous iterations of the belief propagation. In the latter case, the similarity of two domains is based on the overlap in hosts that contacted them, time difference between connections by the same host, and proximity of their IP addresses. These features, combined with properties of the domain itself (domain age, registration validity, popularity of user-agent string and web referer information), are used in a regression model to compute a *similarity score* for the domain relative to the set of domains already labeled suspicious during belief propagation. More details about domain similarity is provided in §IV-D.

#### D. Detecting C&C communication

As discussed in §II communication to the C&C center happens in many campaigns on a regular basis to allow the attacker backdoor access into the compromised system. Such communications are automatically generated by a malicious process on the victim’s machine, and exhibit certain regularity in connection timing patterns. By contrast, normal user activities are largely variable.

We leverage these insights to build a novel detector for automated communication that compares the inter-connection histogram of the communication between a host and a domain to that of a periodic (regular) distribution. The communication between a host and a domain is labeled “automated” if the statistical distance between the two histograms is below a certain threshold. Compared to other methods for identifying automated connections (e.g., Fast-Fourier transform in BotFinder [37] and autocorrelation in BotSniffer [18]) our method can be tuned for resiliency to outliers and randomization between connections through several parameters. Our detector (like previously proposed timing-based C&C detection methods) will miss communications with large variability in timing patterns, but these methods are not commonly used by attackers [26]. Detecting C&C communication with completely randomized timing patterns (without access to known malware samples) remains an open problem to the community.

However, thousands of legitimate requests have regular timing patterns as well (due to site refreshes or automatic

updates). Restricting our focus to *rare* domains significantly reduces the number of considered domains, but we need to leverage additional features to identify the truly suspicious automated connections. Some of these features have been used in previous work for identifying generic malicious activities (e.g., domain age and registration validity extracted from WHOIS data), and some are tailored to an enterprise environment (e.g., popularity of user-agent strings within that network and web referer information in the HTTP traffic).

Combining these features, we train a linear regression model to output a score for each domain that is detected to receive automated, periodic, communications. More specifically, the score for a domain is a weighted sum of the features, where the weights are determined by the regression model during training. Domains with score higher than a threshold (determined based on tradeoffs between accuracy and coverage) are considered potential C&C domains.

#### E. Putting it all together

Our system for detecting early-stage enterprise infection consists of two main phases: training (during a one-month bootstrapping period) and operation (daily after the training period). An overview diagram is presented in Figure 1.

**Training.** The training period is specific to each organization and its role is to create a benchmark of normal activity of enterprise hosts. It consists of several steps.

- (1) *Data normalization and reduction:* The first stage processes the raw log data (either HTTP or DNS logs) used for training and applies normalization and reduction techniques.
- (2) *Profiling:* Starting from normalized data, the system profiles the activity of internal hosts. It builds histories of external destinations visited by internal hosts as well as user-agent (UA) strings used in HTTP requests (when available). These histories are maintained and incrementally updated during the operation stage when new data is available.
- (3) *Customizing the C&C detector:* The detector of C&C communication is customized to the characteristics of the enterprise. Enterprise-specific features (e.g., rare destinations, popularity of user-agent strings used in a connection, web referer information) are used in combination with other generic features utilized in previous work (automated connections, domain age and validity). A regression model is trained to determine the feature weights for the particular enterprise and the threshold above which a domain is flagged as C&C.
- (4) *Customizing the domain similarity score:* The domain similarity score used during belief propagation is also customized to the enterprise during the training phase. The weights of the features used for determining domain similarity scores during belief propagation, as well as the score threshold, are also chosen by a regression model.

**Operation.** After the initial training period, the system enters into daily operation mode. Several stages are performed daily:

- (1) *Data normalization and reduction:* The system processes new data for that day, normalizes it and performs data reduction.
- (2) *Profile comparison and update:* New data is compared with historical profiles, and rare destinations, as well as rare UAs

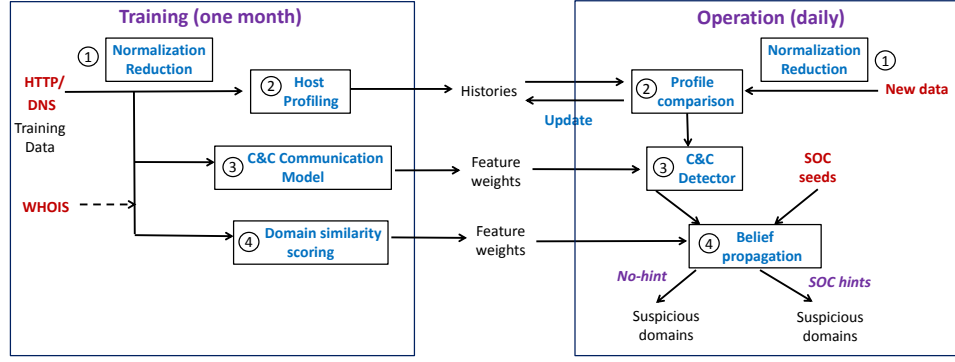


Fig. 1: Overview of training and operation stages in our system for detecting enterprise infection. Training stage is on the left and operation on the right. Input data is shown in red, processing steps in blue and various outputs in black.

(used by a small number of hosts) are identified. Histories of external destinations and UAs are updated with new data, so that changes in normal behavior are captured in the profiles. (3) *C&C detector*: The C&C detector is run daily, and scores of automated domains are computed with weights determined during training. Automated domains with scores above a threshold are labeled as potential C&C domains. (4) *Belief propagation*: The belief propagation algorithm is run in two modes (with or without hints). The output is an ordered list of suspicious domains presented to SOC for further investigation.

#### IV. SYSTEM DETAILS

After providing an overview of our system for detecting early-stage enterprise infection, we give here more technical details of our methods.

##### A. Data Normalization and Reduction

**LANL dataset.** LANL released a dataset comprised of anonymized DNS traffic collected over two months (February and March 2013) from their internal network [15]. The entire LANL dataset consists of 3.81 billion DNS queries and 3.89 billion DNS responses, amounting to 1.15 TB. To allow efficient analysis, we employ a number of data reduction techniques. We first restrict our analysis only to A records, as they record the queries to domain names and their responses (IP addresses) and information in other records (e.g., TXT) is redacted and thus not useful. This step prunes 30.4% of DNS records on average per day. We also filter out queries for internal LANL resources (as our focus is on detecting suspicious external communications), and queries initiated by internal servers (since we aim at detecting compromised hosts).

**AC dataset.** The AC dataset consists of web proxies logs generated at the border of a large enterprise over a period of two months (January and February 2014). Analyzing the AC dataset proved difficult due to its large scale, as well as inconsistent information. There are on average 662GB of log data generated daily, resulting in a total of 38.14TB of data over the two months of our investigation. This dataset is 33 times larger than the LANL dataset, and much richer in

information. Compared to LANL data in which all timestamps are in the same time zone and IP addresses are statically assigned, the AC dataset has some inconsistencies due to multiple time zones (as the collection devices are in different geographical locations) and DHCP and VPN assignment for most of the IP address space.

We omit here a detailed description of our normalization procedure, but we converted all timestamps into UTC and DHCP and VPN IP addresses to hostnames (by parsing the DHCP and VPN logs collected by the organization). After normalization, we extract the *timestamp*, *hostname*, *destination domain*, *destination IP*, *user-agent string*, *web referrer* and *HTTP status code* fields for our analysis. We do not consider destinations that are IP addresses.

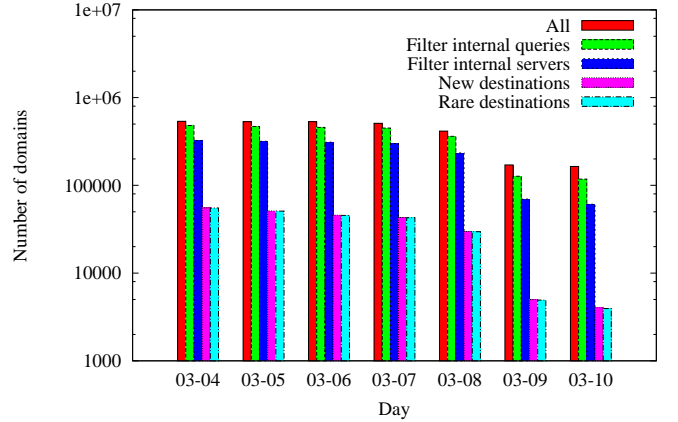


Fig. 2: The number of domains encountered daily in LANL after data reduction for the first week of March.

**Rare destinations.** For both datasets, we use the first month of data for profiling and building a history of external destinations visited by internal hosts. We determine for each day of the second month a list of new destinations not visited before by any internal host. The rationale is that attackers tend to use new domains under their control not visited previously by internal hosts, and benign hosts (the large majority) are unlikely to visit malicious sites contacted by compromised hosts.

For profiling external destinations, we first “fold” the domain names to second-level (e.g., `news.nbc.com` is folded to `nbc.com`), assuming that this captures the entity or organization responsible for the domain. Since domain names are anonymized in the LANL dataset, we conservatively fold to third-level domains (as we don’t have information on top-level domains). We then maintain a history of (folded) external destinations queried by internal hosts, updated at the end of each day to include all new domains from that day. A domain is considered to be *new* on a particular day if it has not been contacted before, and *unpopular* if it has been queried by less than a certain threshold of distinct hosts in a single day. We set the threshold at 10 hosts based on discussion with security professionals at a large organization. These *rare destinations* are the starting point for our investigation.

Following the steps detailed above, we greatly reduce the size of both datasets. The number of domains after each reduction step in LANL for one week in March is shown in Figure 2. On average, while the full dataset contains queries from almost 80,000 hosts to more than 400,000 domains per day, in the reduced dataset only 3,369 hosts querying 31,582 domains are included daily on average. In the AC dataset starting from 120K hosts and 600K domains in the original dataset, after data reduction we encounter on average 20K hosts and 59K rare domains daily.

### B. Belief Propagation Algorithm

The belief propagation algorithm can be applied in two modes: with hints of compromised hosts provided by SOC, or without hints. In the first case we use as seed a list of compromised hosts investigated and confirmed by security analysts in SOC and optionally a list of malicious domains. In the latter case the C&C communication detector is run first to identify a set of potential C&C domains and hosts contacting them. These are given as seeds to the belief propagation algorithm. The algorithm is run daily in both modes and it detects malicious domains and compromised hosts that are likely part of the same campaign with the provided seeds.

Algorithm 1 presents the pseudocode of the belief propagation algorithm. The algorithm is initialized with the set of compromised hosts  $\mathcal{H}$ , and set of malicious domains  $\mathcal{M}$  (when available). In the *SOC hints* case,  $\mathcal{H}$  is the set of hint hosts and  $\mathcal{M}$  is the set of malicious domains (if provided). In the *no-hint* case,  $\mathcal{M}$  is the set of potential C&C domains and  $\mathcal{H}$  is the set of hosts contacting them.

The algorithm maintains several variables:  $\mathcal{R}$  the set of rare domains contacted by hosts in  $\mathcal{H}$  and  $\mathcal{N}$  the set of newly labeled malicious domains (in a particular iteration). In each iteration, the algorithm first detects suspicious C&C-like domains among set  $\mathcal{R}$  using function `Detect_C&C` whose exact implementation will be provided next section. Note that in the *no-hint* mode function `Detect_C&C` will not identify additional C&C domains (since they are used for seeding the algorithm and are already included in  $\mathcal{M}$ ). If no suspicious C&C domains are found, the algorithm computes a similarity score for all rare domains in  $\mathcal{R}$  with function `Compute_SimScore`. The domain of maximum score (if

above a certain threshold  $T_s$ ) is included in set  $\mathcal{M}$ . Finally the set of compromised hosts is expanded to include other hosts contacting the newly labeled malicious domain(s). The algorithm iterates until the stopping condition is met: either no new domains are labeled as malicious (due to their scores being below the threshold) or the maximum number of iterations has been reached. The output is an expanded lists of compromised hosts  $\mathcal{H}$  and malicious domains  $\mathcal{M}$ .

---

#### Algorithm 1 [Belief Propagation]

---

```

/*  $\mathcal{H}$   $\leftarrow$  set of seed hosts */
/*  $\mathcal{M}$   $\leftarrow$  set of seed domains */
/* dom_host is a mapping from a domain to set of hosts contacting it */
/* host_rdom is a mapping from a host to set of rare domains visited */
function BELIEF_PROPAGATION( $\mathcal{H}, \mathcal{M}$ ):
     $\mathcal{R} \leftarrow$  set of rare domains contacted by hosts in  $\mathcal{H}$ 
    while stop_condition do
         $\mathcal{N} \leftarrow \Phi$  /* set of newly labeled malicious domains */
        for dom in  $\mathcal{R} \setminus \mathcal{M}$  do
            if Detect_C&C(dom) then
                 $\mathcal{N} \leftarrow \mathcal{N} \cup \{dom\}$ 
                 $\mathcal{R} \leftarrow \mathcal{R} \setminus \{dom\}$ 
        if  $\mathcal{N} = \Phi$  then
            for dom in  $\mathcal{R} \setminus \mathcal{M}$  do
                score[dom]  $\leftarrow$  Compute_SimScore(dom)
                max_score  $\leftarrow$  max[score[dom]]
                max_dom  $\leftarrow$  dom of maximum score
                if max_score  $\geq T_s$  then
                     $\mathcal{N} \leftarrow \mathcal{N} \cup \{dom\}$ 
        if  $\mathcal{N} \neq \Phi$  then
             $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{N}$ 
             $\mathcal{H} \leftarrow \mathcal{H} \cup (\cup_{d \in \mathcal{N}} \text{dom\_host}[d])$ 
             $\mathcal{R} \leftarrow \mathcal{R} \cup (\cup_{h \in \mathcal{H}} \text{host\_rdom}[h])$ 

```

---

### C. Detection of C&C communication

**Dynamic histograms.** We aim at detecting automated connections with fairly regular timing patterns, but be resilient to outliers (for instance large gaps in communication) and small amounts of randomization introduced by attackers between connections.

Initially, we tested a detector for automated connections based on standard deviation (labeling the connections between a host and a domain as automated if the standard deviation of the inter-connection intervals is small), but found out that a single outlier could result in high standard deviation. Our main idea is to generate a histogram of inter-connection intervals between a host and a domain, and compare it to that of a periodic distribution using a known statistical distance. However, the distance metric is highly sensitive to the histogram bin size and alignment. Setting the size to a large value leads to over smoothing, while a small value increases the sampling error. When bins are statically defined, relatively close values might fall under different bins, affecting the distance metric.

We propose a *dynamic histogram* binning method instead. Here we first cluster the inter-connection intervals (denoted  $t_1, \dots, t_m$ ) of successive connections from a host to a domain on a particular day, and then define the bins dynamically from the generated clusters. Let the first interval  $t_1$  be the first cluster “hub.” An interval  $t_i$  is considered as part of a cluster if it is within  $W$  of the cluster hub. Otherwise, a new cluster with hub  $t_i$  is created.  $W$  is a fixed value, and acts as our “bin width.” This dynamic binning method allows us to



accommodate timing randomizations typically introduced by attackers between connections.

Each resulting cluster is considered a bin, its frequency is computed and the resulting histogram is compared to that of the periodic distribution with period equal to the highest-frequency cluster hub. We label the communications between a host and a domain to be automated if their inter-connection histogram is “close” to periodic (i.e., within a threshold  $J_T$ ), as determined by the Jeffrey divergence. For two histograms  $H = [(b_i, h_i)]$  and  $K = [(b_i, k_i)]$ , if  $m_i = (h_i + k_i)/2$  the Jeffrey divergence is defined as [35]:

$$d_J(H, K) = \sum_i (h_i \log \frac{h_i}{m_i} + k_i \log \frac{k_i}{m_i})$$

We choose the Jeffrey divergence metric motivated by the fact that it is “numerically stable and robust to noise and size of histogram bins” [35]. We experimented with other statistical metrics (e.g., L1 distance), but the results were very similar and omit them herein. The bin width  $W$  and distance threshold  $J_T$  parameters control the resiliency of the method to outliers and randomization introduced by attackers. We discuss their selection according to the LANL dataset in §V-B.

**Additional features.** For each rare automated domain visited by an internal host we extract six additional features which will be used for the C&C detector:

*Domain connectivity features:* We consider the number of hosts contacting the domain (NoHosts) called *domain connectivity* and the number of hosts with automated connections to the domain (AutoHosts). The intuition here is that most rare domains (legitimate ones) are contacted by only one host, but those rare domains contacted by multiple hosts are more suspicious as they might indicate multiple compromised hosts under attacker’s control.

*Web connection features:* Based on discussions with security professional at a large organization, web connections with no web referrer information are more suspicious, as they are not part of a user browsing session and might be generated by a process running on the user machine. Typical user browsing sessions include web referer information, but in some cases the web-referer is wiped out (e.g., java script or iframes embedded in web pages). To capture this, we include a feature NoRef denoting the fraction of hosts (among all hosts contacting that domain) that use no web referer.

Additionally, most users have a fairly small number of user-agent strings in their HTTP connections (on average between 7 and 9 per user). Software configurations in an enterprise are more homogenous than in other networks (e.g., university campus), and as such we’d expect that most user-agent strings are employed by a large population of users. With this intuition, the *rare user-agent strings*, those used by a small number of hosts, might indicate unpopular software installed on the user machine which can potentially be associated with suspicious activities. We consider a feature RareUA denoting the fraction of hosts that use no UA or a rare UA when contacting the domain.

To determine the popularity of UA strings, we maintain a history of UAs encountered across time and the hosts using

those UAs. The UA history is built during the training phase for a period of one month and then updated daily based on new ingested data. An UA is considered rare (after the training period of one month) if it is used by less than a threshold of hosts (set at 10 based on SOC recommendation).

*Registration data features:* Attacker-controlled sites tend to use more recently registered domains than legitimate ones [25]. In addition, attackers register their domains for shorter periods of time to minimize their costs in case the campaign is detected and taken down. We query WHOIS information and extract two features: DomAge (number of days since the domain was registered), and DomValidity (number of days until the registration expires).

**Scoring automated domains.** During the training stage, for each rare domain identified as having automated connections, we extract the six features explained above. We also query VirusTotal for each rare automated domain and label it “reported” if at least one anti-virus engine reports it and “legitimate” otherwise. Using the set of automated domains visited in the enterprise for two weeks, we train a linear regression model, implemented using the function `lm` in the R package. The regression model outputs a weight for each feature, as well as the significance of that feature. The final score for each automated domain is a linear combination of feature values weighted by regression coefficients. Finally, based on the model, we select a score threshold  $T_c$  above which domains are labeled as potential C&C domains. We provide results on the regression model and threshold selection in Section VI-A.

During the operation stage, domain scores are computed using the weights of the regression model (built during training). Function `Detect_C&C` in Algorithm 1 returns true for a domain if automated connections are detected to the domain (from at least one host) and the domain score is above the threshold  $T_c$  computed during training.

#### D. Domain similarity

We consider a number of features when computing similarity of a domain  $D$  with a set of domains  $S$  labeled malicious in previous iterations of belief propagation.

*Domain connectivity.* First is the domain connectivity as defined above.

*Timing correlations.* Second, we consider features related to the time when the domain  $D$  was visited by internal hosts. During initial infection stage of a campaign, we suspect that a host visits several domains under the attacker’s control within a relatively short time period (as explained in §II). We thus consider the minimum timing difference between a host visit to domain  $D$  and other malicious domains in set  $S$ . The shorter this interval, the more suspicious  $D$  is.

*IP space proximity.* Third, we consider proximity in IP space between  $D$  and malicious domains in set  $S$ , i.e., if they are in the same IP/16 or IP/24. The intuition here is that attackers host a large number of malicious domains under a small number of IP subnets (i.e., IP/24 or IP/16 subnets) [19], [26].

We provide measurement of the timing and IP proximity features on the LANL dataset in §V-B. In addition, when

Case	Description	Dates	Hint Hosts
1	From one hint host detect the contacted malicious domains.	3/2, 3/3, 3/4, 3/9, 3/10	One per day
2	From a set of hint hosts detect the contacted malicious domains.	3/5, 3/6, 3/7, 3/8, 3/11, 3/12, 3/13	Three or four per day
3	From one hint host detect the contacted malicious domains and other compromised hosts.	3/14, 3/15, 3/17, 3/18, 3/19, 3/20, 3/21	One per day
4	Detect malicious domains and compromised hosts without hint.	3/22	No hints

TABLE I: The four cases in LANL challenge problem.

computing domain scores in belief propagation some of the features introduced for scoring C&C domains (NoRef, RareUA, DomAge, DomValidity) are also used, as they might indicate suspicious activities.

During the training stage we build a linear regression model with eight features in total, using a set of (non-automated) rare domains visited over two weeks. Results and threshold selection are provided in Section VI-A. During the operation stage, function `Compute_SimScore` in Algorithm 1 computes a domain similarity score by using the feature weights found by this model relative to the set of already labeled malicious domains in previous iterations.

## V. EVALUATION ON THE LANL DATASET

We start by describing the four cases in the LANL challenge problem. Then we discuss how we adapted our techniques developed for enterprise infection to detecting the LANL attacks (when less information is available about domain names and connection information). Still, using less features, we are able to demonstrate that our belief propagation framework achieves excellent results on the LANL challenge problem.

### A. The LANL Challenge Problem

The LANL dataset includes attack traces from 20 independent infection campaigns simulated by LANL domain experts. Each simulation is an instance of the initial first-day infection stage of an independent campaign. LANL issued the *APT Infection Discovery Challenge* to the community requesting novel methods for the detection of malicious domains and compromised hosts involved in these attack instances [15]. More specifically, each of the simulated attacks belongs to one of four cases in increasing order of difficulty, described in Table I. Cases 1-3 include “hints” about the identity of one or multiple compromised hosts, while no hint is available for case 4. Answers (i.e., the malicious domains) are also given for each attack for validation.

### B. Parameter selection

When selecting various parameters for our algorithms, we separate the 20 simulated attacks into two equal-size sets, and use one for training (attacks from 3/2, 3/3, 3/4, 3/5, 3/7, 3/12, 3/14, 3/15, 3/17, and 3/18), and the other for testing. We try to include attacks from each case in both training and testing sets, with the only exception of case 4, simulated only on one day (3/22). We deliberately add this most challenging attack (in which no hint is provided) to the

testing set. We use the training set for selecting parameters needed for different components of the algorithm (dynamic histogram method, C&C communication detection, features used for domain scoring, etc.). We show that parameters chosen according to the training set perform well on new data (testing set).

Bin width $W$	Jeffrey distance threshold $J_T$	Malicious pairs in training	Malicious pairs in testing	All automated pairs in testing days
5 seconds	0.0	12	15	7495
	0.034	14	15	8070
	0.06	15	17	8579
	0.35	15	18	34719
10 seconds	0.0	12	16	15611
	0.034	14	16	16224
	0.06	15	18	16803
20 seconds	0.0	12	15	23352
	0.034	14	16	23964
	0.06	15	18	24597

TABLE II: Number of automated malicious (host, domain) pairs in training and testing sets, as well as the number of automated pairs for all days in testing set.

**Thresholds for dynamic histograms.** The dynamic histogram method compares the histogram of inter-connection intervals from successive connections by a host to a domain on a particular day to that of a periodic distribution with the goal of identifying automated communications. As described in §IV-C the method can be configured with two parameters: bin width ( $W$ ) denoting the maximum distance between the bin hub and other intervals in the same bin, and the threshold ( $J_T$ ) denoting the maximum Jeffrey distance between the two histograms. A connection with histogram at distance less than  $J_T$  from the periodic histogram is considered automated. Intuitively, the larger  $W$  and  $J_T$ , the more resilience the method provides against randomization and outliers, but more legitimate connections are labeled automated as well.

To determine the most suitable parameterization, we experiment with 3 different bin widths (5, 10 and 20 seconds) and choose the distance threshold  $J_T$  according to the training set of malicious automated connections. We manually labeled as automated 15 (host, domain) pairs in training set and 18 pairs in testing set corresponding to 18 distinct domains. Table II shows the number of malicious (host, domain) pairs labeled automated (in both training and testing sets), as well as all pairs labeled automated in the testing days for several choices of  $W$  and  $J_T$ . Intuitively, fixing one of the parameters  $W$  or  $J_T$  and increasing the other results in more legitimate domains labeled automated.

For our purposes, we aim at capturing all malicious pairs in the training and testing sets, while labeling fewest legitimate connections automated. For the 10 and 20-second bin width, the threshold capturing all 33 malicious pairs is 0.06, while for the 5-second bin width we need to increase the threshold at 0.35. This larger threshold has the effect of increasing the number of legitimate pairs labeled automated. Based on these results, we choose a bin size of 10 seconds and a threshold distance of 0.06 to achieve our desired properties.

**Features.** Since domain names in the LANL dataset are anonymized and the data contains only DNS requests, we have

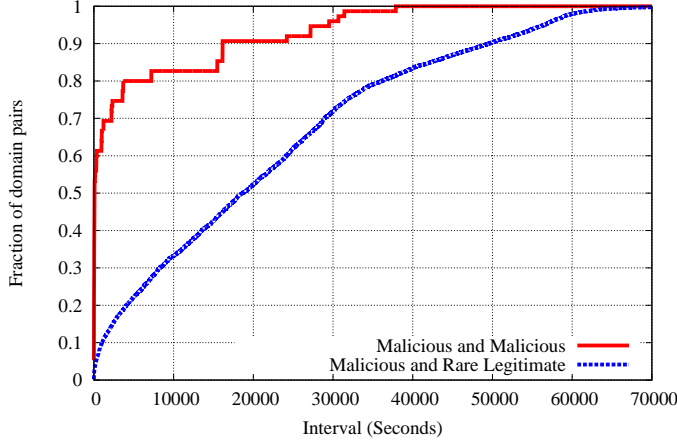


Fig. 3: The CDFs between first connection to two malicious domains and a malicious and legitimate domain by a host.

access to a smaller number of features than in the enterprise case. Features related to domain registration (domain age and registration validity), and features extracted from HTTP connections are not available in LANL.

For detecting C&C communication, there are thousands of automated domains daily (up to 5239). Restricting to rare domains is beneficial in reducing the number of automated domains by a factor of more than 100, but we still observe hundreds of rare automated domains. To identify C&C communications among this set, we combine multiple features that are available in this dataset, in particular domain connectivity and similarity in timing patterns across hosts. The C&C detector for LANL is very simple: we consider an automated domain as potential C&C if there are at least two distinct hosts communicating with the domain at similar time periods (within 10 seconds). This heuristic works well because in the LANL simulations there are always multiple infected hosts in every campaign. However, in our most general C&C detector for the enterprise case, we consider domain connectivity as a feature which can be combined with other features extracted from registration data and HTTP connection. Our general method can detect C&C domains contacted by a single host.

Using the LANL simulated attacks, we’d like to measure the relevance of the timing and IP space similarity features among malicious domains. For compromised hosts in the training set, we extract the timestamp of their first connection to every rare domain visited. We plot in Figure 3 CDFs of the distributions of the time difference between visits to malicious domains and a legitimate and malicious domain by the same host. The graph confirms that connection intervals between two malicious domains are much shorter than between a malicious and a legitimate domain. For example, 56% of visits to two malicious domains happen at intervals smaller than 160 seconds, while only 3.8% of malicious-to-legitimate connection intervals are below this threshold (similar results are observed on testing dataset).

Next we measure similarity in IP space for malicious

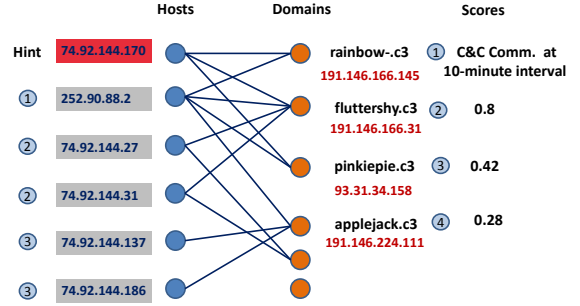


Fig. 4: Application of belief propagation to the 3/19 campaign.

and legitimate domains in the training set. We found that 7 malicious domain pairs are in the same IP/24 subnet, while 18 share an IP/16 subnet. We observed few cases of legitimate domains residing in the same subnet with malicious ones. With the exception of 3/7, when more than 2000 pairs of malicious and legitimate domains share the same IP24 or IP16 subnet (due to a single malicious domain belonging to a popular service), the rest of days we observe 20 pairs in the same IP24 subnet and 155 pairs in the same IP16 subnet. We thus use both IP/16 and IP/24 space features in detection but with different weights.

**Domain similarity scores.** In a particular iteration of belief propagation a domain  $D$  is compared to the set of domains  $S$  already labeled malicious in previous iterations. The domain score is computed as a function of three components: domain connectivity, timing correlation with a known malicious domain in  $S$  (value 1 if the domain is contacted close in time to a malicious domain and 0 otherwise), proximity in the IP space with malicious domains in  $\mathcal{M}$  (value 2 if same /24 subnet with a malicious domain, 1 if same /16 subnet with a malicious domain and 0 otherwise). Each of these components increases the domain’s score. While in our general framework, we proposed a linear regression model for computing domain scores (see §IV-C and §IV-D), we cannot apply that technique here due to limited samples of training data. Instead, we choose a simple additive function that computes the score as the sum of the three components above and then normalizes it. This performs well in both cases 3 and 4 (based on the training set we set the domain score threshold  $T_s$  at 0.25).

### C. Results

We omit here description of the first two cases due to space limitations.

**Starting from a hint host (case 3).** We ran the belief propagation algorithm for a maximum of five iterations starting from the provided hint host, but we stop the algorithm if the maximum domain score is below the threshold  $T_s$ .

An example of applying the algorithm to detecting the campaign on 3/19 in the LANL dataset is given in Figure 4. Starting from hint host 74.92.144.170, in the first iteration

C&C communication to domain `rainbow-.c3` is detected at 10-minute average intervals. The domain is labeled malicious and the second host (252.90.88.2) compromised. In the following three iterations, three rare domains of maximum score are labeled malicious based on similarity with previously labeled malicious domains (their scores are given on the right). The algorithm stops at the fifth iteration when the maximum score is below a threshold, at which point all labeled domains are confirmed malicious.

The algorithm detects all 12 labeled malicious domains in training days (with no false positives) and all 12 labeled malicious domains in testing days at the cost of one false positive for 3/21.

**No hints (case 4).** In the most challenging case no hint on compromised hosts are available to seed the belief propagation algorithm. We thus identify first C&C communication and seed the belief propagation algorithm with the C&C domains. Interestingly, the same algorithm from case 3 delivered very good results on case 4, where we did not have an opportunity for training (case 4 was simulated only on a single day). The five domains identified by belief propagation were confirmed malicious (by the simulation answer), and the algorithm did not have any false positives.

**Summary.** The summary of our results on the four cases of the LANL challenge are given in Table III. We define several metrics of interest: *true detection rate* (TDR) is the fraction of true positives and *false detection rate* (FDR) is the fraction of false positives among all detected domains; and *false negative rate* (FNR) is the fraction of malicious domains labeled as legitimate by our detector. Overall, we achieve TDR of 98.33% (97.06% on the testing set), with FDR of 1.67% (2.94% on testing set) and FNR of 6.35% (2.94% on the testing set).

Case	True Positives		False Positives		False Negatives	
	Training	Testing	Training	Testing	Training	Testing
Case 1	6	4	0	0	2	0
Case 2	8	12	0	0	1	1
Case 3	12	12	0	1	0	0
Case 4	-	5	-	0	-	0
Total	26	33	0	1	3	1

TABLE III: Results on LANL challenge.

## VI. EVALUATION ON ENTERPRISE DATA

We implemented a fully operational system running in production starting from January 1 2014 to process the web proxies logs from the AC dataset. We use the data collected in January for training various components of the system (e.g., the C&C detector, the domain scoring module, etc.) and profiling external destinations and user-agent strings used by enterprise hosts in HTTP communication. Starting from February 1 the system enters into the daily operation mode, in which it processes new web proxies logs, applies normalization and reduction techniques, compares the data with the profiles (which are also updated) and applies our detection techniques. First C&C communication is identified, and domains labeled as C&C are given as seeds to belief propagation. Second, domains confirmed by SOC (if available) and hosts contacting them are provided as seeds in belief propagation. The detection

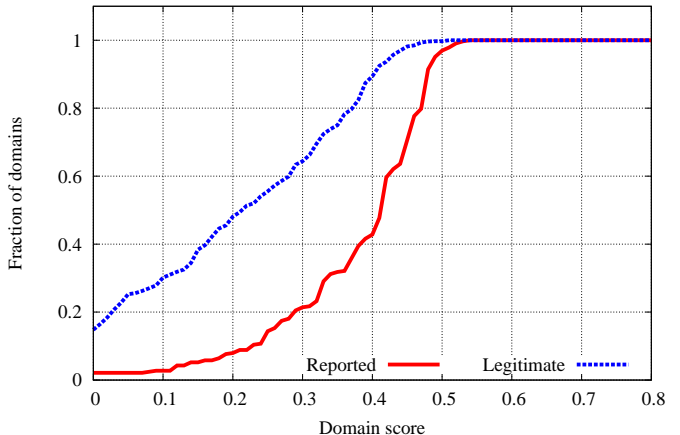


Fig. 5: CDFs of automated reported/legitimate domain scores.

results are thoroughly validated through a combinations of tools and manual analysis. The system is configurable with different parameters (e.g., scoring thresholds, number of iterations in belief propagation, etc.) according to the SOC’s processing capacity. We present the parameter selection, our validation methodology and the results in both modes of operation.

### A. Parameter selection

**Scoring automated domains.** The parameters of the dynamic histogram method selected in §V-B result in 841 distinct automated domains for the month of February in the AC dataset. We query VirusTotal for each rare automated domain and label it “reported” if at least one anti-virus engine reports it and “legitimate” otherwise (144 domains were reported by VirusTotal). We thus generate a large set of labeled automated domains and use it to learn the relevance of each feature in identifying suspicious C&C domains.

In total, we observe 2888 automated (host, domain) pairs in the month of February for which we can extract all features described in §IV-C. We divide the labeled set of automated domains into a training set covering the first two weeks of February, and testing set for the last two weeks. We construct a linear regression model on the training set, which outputs a coefficient for each feature, as well as the significance of that feature. Among all features considered, the only one with low significance was AutoHosts, which we believe is highly correlated with NoHosts and thus omit it. The most relevant features found by the model are DomAge and RareUA. DomAge is the only one negatively correlated with reported domains (as they are in general more recently registered than legitimate ones), but all other features are positively correlated.

The graph in Figure 5 shows the difference between the scores of automated domains reported by VirusTotal and legitimate ones on the training set. We observe that reported domains have higher scores than legitimate ones. For instance, selecting a threshold of 0.4 for labeling an automated domain suspicious results in 57.18% true detection rate and 10.59%

false positive rate on the training set, and 54.95% true detection rate and 11.52% false positives on the testing set. We emphasize that our final goal is not identifying **all** automated domains reported by VirusTotal, but rather a significant fraction that allow us to bootstrap the belief propagation algorithm and find new suspicious domains not yet detected by current anti-virus technologies.

**Domain similarity.** To obtain a list of (non-automated) rare domains and their features, we start from a set of compromised hosts (those contacting the C&C domains confirmed by VirusTotal). We include each rare domain contacted by at least one host in this set, extract its features, query VirusTotal to get an indication of its status, and divide the data into training and testing set, covering the first and last two weeks of February, respectively.

We apply again linear regression on the training set to determine feature weights and significance. Among the eight considered features described in §IV-D, the only one with low significance was IP16, as we believe it's highly correlated with IP24. The most relevant features identified by the model are RareUA, DomInterval, IP24 and DomAge. We still observe a difference in the score CDFs for reported and legitimate domains (we omit the plot due to space limitations).

### B. Validation methodology

We ran the detector in both modes, *SOC hints* and *no-hints*. In the *SOC hints* mode we use domains from the IOC list provided by SOC as seeds. The domains output by our detector in both modes (not considering the seeds provided by SOC) were validated as follows. We first query VirusTotal and the IOC domain list to verify their status (three months after they were detected – to allow anti-virus and blacklists to catch up). If the domain is alerted upon by at least one scanner used by VirusTotal or it's an IOC we consider it *known malicious*. For other domains, we collect additional information and hand them to a security analyst for manual investigation. Specifically, we retrieve the associated URLs from the log data and crawl them to examine the responses. The URLs are also manually submitted to McAfee SiteAdvisor. Based on the URLs, the response to our crawler and the result from SiteAdvisor, we classify the remaining domains into four categories: *new malicious* (e.g., same URL patterns as known malicious domains, returning malicious content or flagged by SiteAdvisor), *suspicious* (not resolvable when crawled, parked or having some questionable activities), *legitimate* (no suspicious behavior or code observed) and *unknown* (504 HTTP response code, a sign of server error). Since we only have a few unknowns (6 in total), we remove them from the final results. When reporting our results we use several metrics of interest: TDR and FDR defined in §V-C, and *new-discovery rate* (NDR) defined as the percentage of new malicious and suspicious domains detected. Here TDR is the percentage of both known and new malicious and suspicious domains among all detected domains, and  $FDR = 1 - TDR$ .

### C. Results for the no-hint case

We first evaluate the effectiveness of our detector in the *no-hints* mode. We compute scores for all automated domains

encountered daily. For domains whose WHOIS information can not be parsed default values for DomAge and DomValidity are set at average values across all automated domains. As the first step, we vary the threshold for labeling automated connections from 0.4 to 0.48 and present results for domains detected as C&C domains (with score above the threshold) in Figure 6(a). The results demonstrate that as we increase the threshold on automated domain scores from 0.4 to 0.48 the number of domains labeled as C&C drops from 114 to 19, while accuracy increases (TDR increases from 85.08% to 94.7%). Though FDR is higher for threshold 0.4, more malicious domains (including 23 new ones not known to VirusTotal or SOC) are detected. They can be used to seed the belief propagation stage, and therefore we fix the threshold at 0.4 to evaluate the overall effectiveness of belief propagation in this mode.

Next, we vary the threshold for domain scoring in belief propagation from 0.33 to 0.85 and the result (Figure 6(b)) shows that the number of all detected domains varies from 265 to 114, with TDR ranging from 76.2% to 85.1%. Altogether in the most challenging case (when no hint is available), we detect 202 malicious and suspicious domains in February, associated with 945 hosts. Though the majority of the detected domains are already alarmed by SOC and VirusTotal (132 for threshold 0.33), only 13 are reported in the IOC list and the remaining ones are unknown to the enterprise. More interestingly, we identified many new malicious and suspicious domains not known to the community (a total of 70 new domains for threshold 0.33 resulting in an NDR of 26.4%). This result suggests that our detector could greatly complement existing security tools by discovering new suspicious activities. Its main advantage is that it has the ability to detect new campaigns without traces of known malicious behavior.

We thoroughly examined the domains labeled as new malicious and suspicious and found several prominent and interesting clusters. Among the new malicious domains, we found 5 domains hosting URLs with the same pattern `/logo.gif?` later confirmed by the SOC as related to Salty worm. We also found 15 domains with the same URL pattern reported by VirusTotal. Moreover, we identified a cluster of 10 DGA domains with none of them reported by VirusTotal and SOC, demonstrating our detector's ability in capturing new malicious campaigns. All the malicious domains are under the TLD `.info` and their names have 4 or 5 characters (e.g., `mgwg.info`). 9 out of the 10 domains hosts URLs with pattern `/tan2.html` and visiting them will be redirected to the remaining domain `1.tv990.info`.

We labeled legitimate a set of 63 domains belonging to categories like Ad-network, Gaming, Toolbar and Torrent Tracker. They are captured by our detector because they exhibit suspicious features, like automated connections or are registered recently. Though they do not pose serious harm to the enterprise, some of them are policy violations (e.g., Gaming, Torrent Tracker). We labeled them legitimate since we did not discover any suspicious activities, but we believe these domains still need to be vetted.



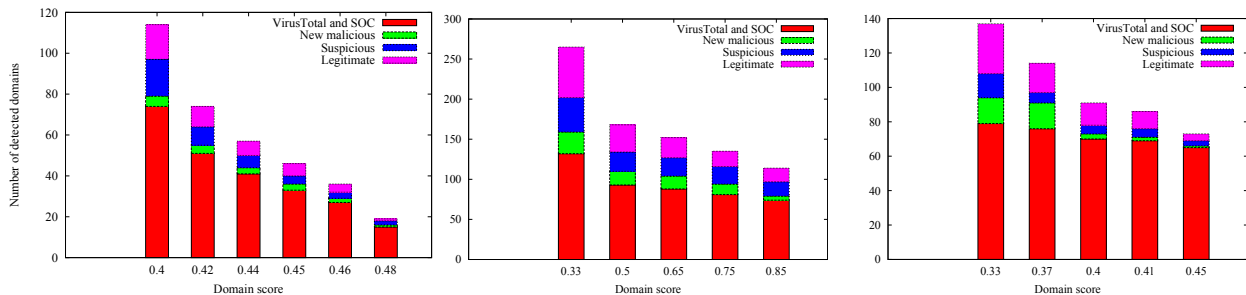


Fig. 6: Statistics on detected domains. (a) C&C communication. (b) No hints. (c) SOC hints.

#### D. Results for the SOC hints case

We also present results in Figure 6(c) for the belief propagation algorithm in *SOC hints* mode seeded with 28 IOC domains. We do not include the seed domains in the results. We set the domain score threshold at 0.4 for automated domains and vary the similarity score between 0.33 and 0.45. In total, we detect between 137 domains (at threshold 0.33) to 73 domains (at threshold 0.45), with TDR ranging from 78.8% to 94.6%. Among the 137 detected domains, 108 turn out to be malicious (either known or new) and suspicious, which is about four times larger than the set used for seeding.

Among the 108 malicious and suspicious domains, 79 domains are confirmed by SOC or VirusTotal, leaving 29 domains as our new findings. We inspect the new findings and identify an interesting group of domains generated through DGA. This group consists of 10 domains under TLD `.info` and the name for each domain has 20 characters (e.g., `f0371288e0a20a541328.info`). Surprisingly, the registration dates for most of the domains are later than the time when we detected them. For example, one domain is detected on 2014/02/13 but registered on 2014/02/18. Attackers use DGA domains to increase the robustness of their C&C centers against take-downs, and they only register a portion of the domains to reduce the cost. Our detector is able to detect the malicious domains before registration and obtain an advantage in the arm-race.

Finally, we compare the results of the two modes of operation. Only 21 domains are detected in both modes, which is a small portion compared to 202 and 108 malicious and suspicious domains detected separately. When deployed by the enterprise, we suggest our detector configured to run in both modes, in order to have better coverage. We present two case studies for both modes of operation in the Appendix. As we have shown, starting from a seed of known malicious domains or hosts, the algorithm in *SOC hints* mode can identify suspicious domains with high accuracy. The *no-hint* case has the unique capability of identifying new unknown attack campaigns, especially C&C communications of these campaigns (even when only a single host is compromised). We recommend that C&C detected domains are first vetted by the SOC and then the algorithm can be used in the *SOC hints* mode for those confirmed malicious domains.

Both variants include configurable thresholds for scoring automated and non-automated domains. These thresholds can be chosen by the SOC according to the capacity of the

team performing manual investigation, and various tradeoffs between accuracy and larger coverage.

## VII. RELATED WORK

Our work focuses on detecting early-stage infections within enterprise perimeters, including communications related to malware delivery and C&C. There has been a large body of work in this area, but to the best of our knowledge, we are the first to exploit the relationship between malicious domains associated with the same attack campaign, and to detect them by a graph-theoretic framework based on belief propagation. We describe here related work in the literature.

**Detection of C&C communication.** Some of the previous work detecting C&C domains in botnets require malware samples as input to detect connections with similar patterns (e.g., BotFinder[37], Jackstraws[22]). Anomaly-based botnet detection systems (e.g., BotMiner[16], BotSniffer[18] and TAMD [41]) typically detect clusters of multiple synchronized hosts infected by the same malware. In contrast to these, our approach does not require malware samples and can detect a single compromised host contacting the C&C server.

DISCLOSURE [8] identifies C&C traffic using features extracted from NetFlow records but incorporates external intelligence sources to reduce false positives. Our C&C detector is different in that it leverages enterprise-specific features extracted from HTTP connections. From that perspective, ExecScent [28] is close to our work in detecting C&C communications in large enterprise network. However, ExecScent needs malware samples to extract templates representing malicious C&C connections. The templates are adapted to a specific enterprise considering the popularity of different features (URL patterns, user-agent strings, etc.). Our work complements ExecScent in detecting new unknown malware that can be provided as input to the template generation module.

**Detection of malware delivery.** Nazca [21] analyzes web requests from ISP networks to identify traffic related to malware delivery and unveils malicious distribution networks. CAMP [32] determines reputation of binary downloads in the browser and predicts malicious activities. BotHunter [17] identifies sequences of events during infection, as observed from a network perimeter. Our approach does not depend on specific events occurring during infection (which can be changed easily or may not be observable), but more focused on detecting related malicious domains and compromised hosts.

**Detection of malicious domains.** Domains used in malicious activities are backed by highly resilient infrastructures to deal with takedowns or blacklisting, and hence exhibit unique characteristics distinct from benign sites. Another branch of work detects domains involved in malicious activities by patterns observed in DNS traffic (e.g., EXPOSURE [9], Notos [5], Kopis [6], and Antonakakis et al. [7]). Paxson et al. [30] detect malicious communication established through DNS tunnels. Carter et al. [10] use community detection for identifying highly localized malicious domains in the IP space.

**Anomaly detection in enterprise network.** Beehive [40] is an unsupervised system identifying general anomalies in an enterprise setting including policy violations and malware distribution. Our work is specifically targeting enterprise infections which pose high risk and potential financial loss.

**Targeted attacks.** The threats in cyberspace keep evolving and more sophisticated attacks recently emerged. Some targeted attacks (APT) are well-funded, carefully orchestrated and persist in the victim environments for years before detection.

Detecting targeted attacks in general is a very challenging task. These attacks are usually very stealthy and able to evade existing defenses [3]. However during the automated infection stage many campaigns (e.g., Shady RAT [20], Mirage [12], APT1 [26]) exhibit similar infection patterns. Recent studies have shown that even though in theory APTs could be arbitrarily sophisticated, in practice goal-oriented attackers use relatively low levels of sophistication [38], [27], [24]. We leverage some common patterns observed during the infection stage to build a detector tailored to an enterprise. Our detection result on the LANL's APT infection discovery challenge indicates that our techniques have potential in detecting infections originated from targeted attacks.

### VIII. LIMITATIONS AND DISCUSSION

Attackers could in principle use a number of techniques to evade our detectors. For instance, they may communicate through different channels than HTTP/HTTPS but other ports are commonly blocked at enterprise borders. Attackers could also compromise popular web sites for delivering malicious payloads or C&C communications. However, this is not broadly observed. The analysis of traffic to popular sites require different models and assumptions, due to the amount of (benign) noise. As another evasion technique, attackers can randomize timing patterns to C&C servers, but according to published reports ([26]) this is uncommon. Our dynamic histogram method is resilient against small amounts of randomization introduced by attackers. Detecting C&C communication with completely randomized timing patterns (without having access to malware samples and without correlating activity from multiple hosts) is a challenging problem to the community. Nevertheless, we believe that the infection patterns that we detect are quite prevalent in many attacks.

The approach we proposed is meant to complement the existing tools rather than replace them. The results from §VI demonstrate that our belief propagation algorithm in both variants (*SOC hints* and *no-hint*) detects new suspicious activities overlooked by deployed defense mechanisms. These include

both domains associated with existing malware campaigns (and identified by VirusTotal), but with new presence in the enterprise of our study, as well as entirely new malware campaigns (not yet detected by anti-virus technologies). Since our methods are focused on detecting the initial infection stages of a campaign it is difficult to determine how many of these suspicious activities are related to more advanced attacks, and how many are mainstream malware variants. We believe that monitoring activity to these suspicious domains over longer periods of time, as well as correlating with information from other data sources will answer this question, and we leave this as an interesting avenue for future work.

### REFERENCES

- [1] Hackers in China attacked The Times for last 4 months. <http://www.nytimes.com/2013/01/31/technology/chinese-hackers-infiltrate-new-york-times/>, 2013.
- [2] Target's data breach: The commercialization of APT. <http://www.securityweek.com/targets-data-breach-commercialization-apt/>, 2014.
- [3] Verizon 2014 data breach investigations report. <http://www.verizonenterprise.com/DBIR/2014/>, 2014.
- [4] Alexa. AWS — Alexa Top Sites - Up-to-date lists of the top sites on the web. <http://aws.amazon.com/alexa-top-sites/>, 2014.
- [5] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster. Building a dynamic reputation system for DNS. In *Proc. 19th USENIX Security Symposium*, 2010.
- [6] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, II, and D. Dagon. Detecting malware domains at the upper DNS hierarchy. In *Proc. 20th USENIX Security Symposium*, 2011.
- [7] M. Antonakakis, R. Perdisci, Y. Nadjai, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon. From throw-away traffic to bots: Detecting the rise of DGA-based malware. In *Proc. 21st USENIX Security Symposium*, 2012.
- [8] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, and C. Kruegel. Disclosure: Detecting botnet Command-and-Control servers through large-scale NetFlow analysis. In *Proc. 28th Annual Computer Security Applications Conference, ACSAC*, 2012.
- [9] L. Bilge, E. Kirda, K. Christopher, and M. Balduzzi. EXPOSURE: Finding malicious domains using passive DNS analysis. In *Proc. 18th Symposium on Network and Distributed System Security, NDSS*, 2011.
- [10] K. M. Carter, N. Idika, and W. W. Streilein. Probabilistic threat propagation for network security. *IEEE Transactions on Information Forensics and Security*, 9:–, 2014.
- [11] Command Five Pty Ltd. Command and control in the fifth domain. [http://www.commandfive.com/papers/C5\\_APT\\_C2InTheFifthDomain.pdf](http://www.commandfive.com/papers/C5_APT_C2InTheFifthDomain.pdf), 2012.
- [12] Dell SecureWorks. The Mirage campaign. <http://www.secureworks.com/cyber-threat-intelligence/threats/the-mirage-campaign/>, 2012.
- [13] Dell SecureWorks. Top banking botnets of 2013. <http://www.secureworks.com/cyber-threat-intelligence/threats/top-banking-botnets-of-2014/>, 2014.
- [14] N. Falliere, L. O. Murchu, and E. Chien. W32.Stuxnet dossier. [http://www.symantec.com/security\\_response/whitepapers.jsp](http://www.symantec.com/security_response/whitepapers.jsp), 2011.
- [15] P. Ferrell. APT infection discovery using DNS data. C3E Challenge Problem. <http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-13-23109>, 2013.
- [16] G. Gu, R. Perdisci, J. Zhang, and W. Lee. BotMiner: Clustering analysis of network traffic for protocol and structure-independent botnet detection. In *Proc. 17th USENIX Security Symposium*, 2008.
- [17] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee. BotHunter: Detecting malware infection through IDS-driven dialog correlation. In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, SS'07, pages 12:1–12:16, Berkeley, CA, USA, 2007. USENIX Association.
- [18] G. Gu, J. Zhang, and W. Lee. BotSniffer: Detecting botnet command and control channels in network traffic. In *Proc. 15th Network and Distributed System Security Symposium, NDSS*, 2008.
- [19] S. Hao, N. Feamster, and R. Pandrangi. Monitoring the initial DNS behavior of malicious domains. In *Proc. ACM Internet Measurement Conference, IMC '11*, 2011.

- [20] Hon Lau. The truth behind the Shady RAT. <http://www.symantec.com/connect/blogs/truth-behind-shady-rat>, 2011.
- [21] L. Invernizzi, S. Miskovic, R. Torres, S. Saha, S.-J. Lee, C. Kruegel, and G. Vigna. Nazca: Detecting malware distribution in large-scale networks. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS '14)*, Feb 2014.
- [22] G. Jacob, R. Hund, C. Kruegel, and T. Holz. Jackstraws: Picking command and control connections from bot traffic. In *Proc. 20th USENIX Security Symposium*, 2011.
- [23] James Wyke. The ZeroAccess rootkit — Naked Security. <http://nakedsecurity.sophos.com/zeroaccess/>, 2012.
- [24] S. Le Blond, A. Uritesc, C. Gilbert, Z. L. Chua, P. Saxena, and E. Kirda. A look at targeted attacks through the lens of an NGO. In *Proc. 23rd USENIX Security Symposium*, 2014.
- [25] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Beyond blacklists: Learning to detect malicious web sites from suspicious urls. In *Proc. 15th ACM International Conference on Knowledge Discovery and Data Mining*, KDD, 2009.
- [26] MANDIANT. APT1: Exposing one of China's cyber espionage units. Report available from [www.mandiant.com](http://www.mandiant.com), 2013.
- [27] W. Marczak, J. Scott-Railton, M. Marquis-Boire, and V. Paxson. When governments hack opponents: A look at actors and technology. In *Proc. 23rd USENIX Security Symposium*, 2014.
- [28] T. Nelms, R. Perdisci, and M. Ahamad. ExecScent: Mining for new C&C domains in live networks with adaptive control protocol templates. In *Proc. 22nd USENIX Security Symposium*, 2013.
- [29] Panda Security. Annual report PandaLabs - 2013 summary. [http://press.pandasecurity.com/wp-content/uploads/2010/05/PandaLabs-Annual-Report\\_2013.pdf](http://press.pandasecurity.com/wp-content/uploads/2010/05/PandaLabs-Annual-Report_2013.pdf), 2014.
- [30] V. Paxson, M. Christodorescu, M. Javed, J. Rao, R. Sailer, D. Schales, M. P. Stoecklin, K. Thomas, W. Venema, and N. Weaver. Practical comprehensive bounds on surreptitious communication over DNS. In *Proc. 22nd USENIX Security Symposium*, 2013.
- [31] J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Second National Conference on Artificial Intelligence*, 1982.
- [32] M. A. Rajab, L. Ballard, N. Lutz, P. Mavrommatis, and N. Provos. CAMP: content-agnostic malware protection. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS '13)*, Feb 2013.
- [33] U. Rivner. Anatomy of an attack. [blogs.rsa.com/rivner/anatomy-of-an-attack](http://blogs.rsa.com/rivner/anatomy-of-an-attack), 2011.
- [34] RSA. Stalking the kill chain. <http://www.emc.com/collateral/hardware/solution-overview/h11154-stalking-the-kill-chain-so.pdf>, 2012.
- [35] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.
- [36] G. Stringhini, C. Kruegel, and G. Vigna. Shady Paths: Leveraging surfing crowds to detect malicious web pages. In *Proc. 20th ACM Conference on Computer and Communications Security*, CCS, 2013.
- [37] F. Tegeler, X. Fu, G. Vigna, and C. Kruegel. BotFinder: Finding bots in network traffic without deep packet inspection. In *Proc. 8th International Conference on Emerging Networking Experiments and Technologies*, CoNEXT '12, 2012.
- [38] O. Thonnard, L. Bilge, G. O'Gorman, S. Kiernan, and M. Lee. Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In *Proc. 15th International Symposium on Recent Advances in Intrusion Detection*, RAID, 2012.
- [39] WebSense Security Lab. WebSense 2014 Threat Report. <http://www.websense.com/assets/reports/report-2014-threat-report-en.pdf>, 2014.
- [40] T.-F. Yen, A. Oprea, K. Onarlioglu, T. Leetham, W. Robertson, A. Juels, and E. Kirda. Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks. In *Proceedings of the 29th Annual Computer Security Applications Conference*, ACSAC '13, pages 199–208, New York, NY, USA, 2013. ACM.
- [41] T.-F. Yen and M. K. Reiter. Traffic aggregation for malware detection. In *Proc. Intl. Conf. Detection of Intrusions and Malware, and Vulnerability Assessment*, DIMVA, 2008.

## APPENDIX

Below we report two interesting communities of malicious domains and compromised hosts detected by our techniques.

Figure 7 illustrates an example of a community of malicious domains detected in the *no-hint* mode on 2/13.

We start from detecting suspicious C&C domains (24 domains are identified that day). We pick for illustration one (usteptyshehoaboochu.ru) contacted by three hosts regularly at the same period (120 seconds). This domain is confirmed by VirusTotal and is associated with a malware labeled by Sophos as Troj/Agent-AGLT. Starting from this C&C domain, belief propagation iteratively discovers two other domains confirmed by VirusTotal (likely the delivery stage) and two other hosts connecting to them.

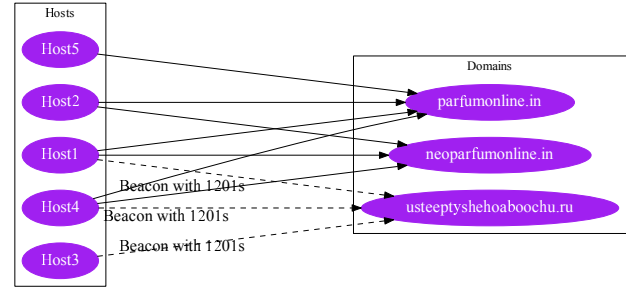


Fig. 7: Example community of compromised hosts and detected domains in *no-hint* mode on 2/13.

Figure 8 illustrates a community of domains detected in the *SOC hints* mode on 2/10. To bootstrap the detection, we use domain xtremesoftnow.ru from the SOC database as a seed. This appears to be a C&C server for Zeus botnet. The domain is accessed by Host 5 which is confirmed infected by the SOC team. Host 5 contacted 7 domains registered under TLD .org. Four of them are confirmed by SOC and also reported by Sophos as contacted by different malwares (Troj Ramdo-B, Troj Ramdo-K, Troj Ramdo-V and Troj Symmi-S). This indicates that the compromised machine downloaded additional malware after the initial compromise. Two of the remaining .org domains are not confirmed by SOC but alarmed by VirusTotal. Interestingly, one domain (uogwoigiuweyccsw.org) has a similar naming pattern with other malicious domains, but is not picked up by either SOC or VirusTotal. This is an example of new discovered domain. The second iteration of belief propagation discovers six additional hosts contacting a similar set of domains as Host 5, indicating that they could be infected with the same malware. Besides, four other suspicious domains are also identified as contacted by these hosts, with three of them confirmed by VirusTotal (including one automated domain) and only one (cdn.tastyreview.com) being legitimate.



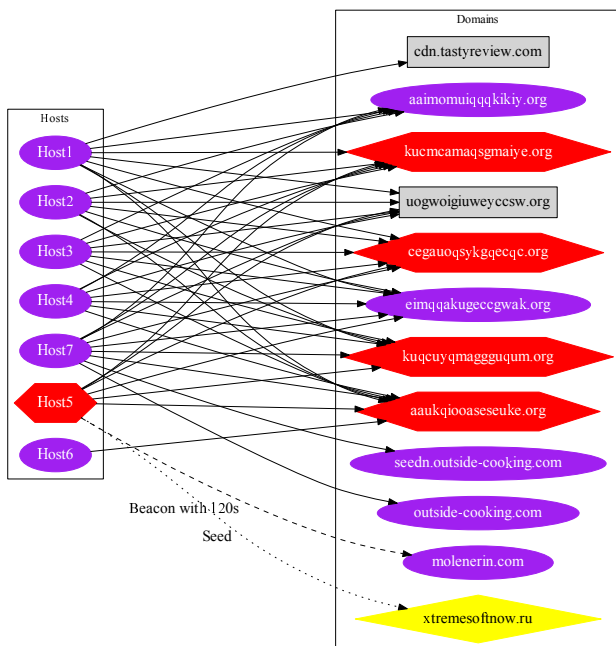


Fig. 8: Example community of compromised hosts and detected domains in *SOC hints* mode on 2/10. The yellow diamond-shape domain is used as seed, purple ellipse-shape domains are detected by VirusTotal and red hexagon-shape domains are confirmed malicious by SOC. The grey rectangle-shape domains are not confirmed by any existing tools at the time of detection. The hostnames are anonymized. Red hexagon-shape hosts are confirmed by SOC and purple ellipse-shape ones are other compromised hosts.